

Published in Technicalities, Mar./Apr. 2007, v. 27, no. 2

The Metadata Layer: Rich Chocolate Value in a Vanilla World

Diane I. Hillmann

January 2007

I'm currently working on a presentation I will give to the Networked Resources and Metadata Interest Group at ALA Midwinter, that I'm calling the Metadata Layer. Sounds like a cake, doesn't it? Aside from the obvious attractiveness of a dessert food analogy, this is an important idea from a number of other viewpoints. I've been working on these ideas with my colleague Jon Phipps, and we're hoping to be able to raise funds to build the application that demonstrates the value we've been talking and writing about for a number of years.

So, what's the problem?

In order to understand why this is such a good idea, we need to take a look at how traditional libraries have lurched into the digital library world. In most cases, they've gone in project-by-project, using grant funds and special project funds to experiment with digitization, search strategies, and display of non-mainstream materials that were the focus of their initial efforts. Years later, they've ended up with a mélange of materials, living in what is best described as a world of digital silos. The content created by their projects is in most cases living in a variety of specially designed and purposed content management systems, and managing these silos and coordinating search across their disparate pockets of metadata (often built as an afterthought) is an increasingly frustrating burden for stressed out library staff.

The sad fact is that even when these projects have reasonably good metadata available, it tends to be managed badly and rarely updated. This sad state of affairs is largely attributable to the expense of coping with metadata within separate content management systems (which, as their name implies, are better at managing content than metadata). This strongly challenges the traditional library commitment to sustainable collections and integrated access, particularly when projects created with grant funds are then expected to be sustained within general library budgets that are already stretched thin. This situation is exacerbated in those libraries now contending with the extreme challenges of coping with mass digitization projects funded in part by Google, Yahoo or Microsoft. The "funded in part" is a critical bit—in most cases the wealthy partner pays for the digital image files and (sometimes) OCR, and the library partners are responsible for storing the image and OCR files, as well as managing the metadata essential for integrating those resources into library services.

I've spoken and written for some time about the challenges of aggregating information harvested from a variety of data providers, but in essence the problems of metadata variety within institutional settings is very similar. The characteristics of both can be stated simply:

- a wide variety of metadata formats
- metadata of varying quality
- limited possibilities to affect the creation and management of the source metadata
- scarce human resources available for maintenance and updating

Libraries in the digital era have gradually reconciled themselves to taking on the role of content provider and publisher, but have not figured out how to integrate those new roles with their traditional role on the

access end. For the most part they have chosen to view the access problem in the context of federated search, where the basic challenge is defined as making search across collections work well enough to present a decent consolidated search result to users. This required building and maintaining a “metasearch” capability that could interact separately with each metadata silo to enable all to be searched at the same time, similarly to how library catalogs are searched using the Z39.50 protocols.

This was all well and good as far as it went, but it didn’t scale particularly well and ignored a rather large elephant in the room: the metadata called upon for the federated search wasn’t getting any better over time, and in fact its usefulness was seriously compromised by the lack of maintenance attention—just the kind of attention that was considered routine for the library’s MARC-based catalog. Names, subjects, and URLs might just as well be cast in stone in this approach, making any improvement in services or attempts to share the metadata with others a risky and potentially embarrassing prospect.

Why a Metadata Layer?

So here’s where I come clean—this isn’t really about layer cakes, or about chocolate. It’s about enabling the use of available metadata to provide richer services over the content that libraries now manage. I still like the layer idea though, because it visually places the “thing” I’m talking about between the various silos that the library must cope with, and the outer world that would like to have access to this metadata for a number of very good reasons. My colleague Jon calls the metadata layer the “metadata washing machine,” which isn’t nearly as sexy a metaphor, but it is nicely descriptive about what really goes on.

So what does this metadata layer look like, and, more importantly, what does it do? Firstly, it’s based on fairly simple harvesting technologies (OAI-PMH), database technologies, and normalization routines not dissimilar to those that bibliographic utilities have been using for years to clean up MARC data for redistribution between libraries. The metadata layer—which we’ll call a Metadata Management System (MMS) for short—harvests the various flavors of metadata from the silos, normalizes it to a standard schema, and serves both the normalized and the original metadata out to other services: search, indexing (Google and others), and other library harvesters.

Now, if this was all the MMS could do for libraries, I’m sure y’all would be underwhelmed. But the normalization and efficiency gains for search, and the ability to expose the metadata to Google is only the beginning of the story. You may recall that I mentioned updating, and this is where the MMS really changes the scenery. But first, a small digression.

One of my current rants (of which I have many) is that libraries look at the problems of metadata as being primarily problems of creation. They tend to see the solution as one of increased standardization—after all, that’s what worked when we all started this data sharing stuff back in the 1970’s and 1980’s. If we could all just agree on some standard metadata schema (many are pushing MODS, but nobody but libraries uses MODS, so don’t get me started on that), our problems with metadata diversity and quality would evaporate. I think that’s a strategy that’s not getting off the ground anytime soon, if ever, and the more we focus on that as our goal, the less likely we are to improve the situation markedly.

So, if the answer to the quality and diversity problems is not standardization at the creation end, where is it? Taking a cue from the 12-steppers, we might start by admitting that we are powerless to change

metadata providers, and that our current approaches to metadata are unmanageable. Given that fact, we need to look at the possibilities for improving what we have, using what we already know about maintaining metadata over time.

In traditional MARC/AACR2 world, we have attempted to improve the precision and recall of the search process by using “access points” (standard forms of names, subject terms, and series). For the most part, we have not taken good advantage of current database management technologies or the insights provided by computer science research, but have instead relied fairly completely on text matching as our only strategy for accomplishing the maintenance of this data. Some of this stubbornness stems from the fact that our vendor-built catalog systems usually display text directly from the records, and lack any more sophisticated methods for managing names and topics. Basic normalization for metadata still relies to some extent on text matching strategies, but doesn’t end there. Ultimately, the transition from text values to URIs for these values is what will enable the provision of the kind of services that will allow us to really improve metadata quality, and for that we’ll need to have a strategy based on an MMS-like layer.

In the UK, the Joint Information Systems Committee (JISC) recently released a draft report on what they’re calling *terminology services*. [1] Their definition of these services hints at the extensive reach of these services in building quality metadata-based applications:

“Terminology Services (TS) are a set of services that present and apply vocabularies, both controlled and uncontrolled, including their member terms, concepts and relationships. This is done for purposes of searching, browsing, discovery, translation, mapping, semantic reasoning, subject indexing and classification, harvesting, alerting etc.”

As a thought experiment, consider for a moment how one would implement terminology services in a metadata environment still based on content management silos, and it should be readily apparent that this is not a solid basis for the future of metadata. With an MMS in place, the way forward is clearly not based on imposing one standard metadata format, but instead reflects a much more realistic view of dealing with what IS available, with tools that allow us to “smarten up” as well as “dumb down.” It’s almost as if we were designing a world where capabilities formerly centralized by bibliographic utilities were now being rebuilt in a decentralized manner. This could not have been contemplated even five years ago, but seems quite possible now.

My vision of this world includes an environment where libraries use an MMS to manage however many content silos they have, and use this capability to develop services that they can use themselves and offer to others (perhaps as combination of free and fee-based). These services might be metadata registries allowing common access to application profiles and crosswalks, terminology services that assist in the substitution of appropriate URIs for textual values, actual crosswalking and vocabulary mapping services (OCLC already has a prototype of these), or metadata augmentation services, that could provide additional topical or classification access to metadata lacking those vital access points.

By the time you see this column, the presentation (with a picture!) will be over, but I plan on making a screencast of the presentation available online afterwards, and will make sure that there are announcements of its URL on various library discussion lists. In the meantime, you might be interested

in reading some of the materials available about our MMS planning. [2] Some additional papers with more detailed explanations of these ideas are also available linked from within that wiki.

[1] Tudhope, Douglas; Koch, Traugott; Heery, Rachel. "Terminology Services and Technology: JISC State of the Art Review." Joint Information Systems Committee (JISC). Available at: http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf

[2] The Metadata Management System wiki. Available at: http://metadataregistry.org/wiki/index.php/Metadata_Management_System